Principal Component Analysis in the Local Differential Privacy Model*

Di Wang† , **Jinhui Xu**
Department of Computer Science and Engineering
State University of New York at Buffalo, NY, USA.
{dwang45,jinhui}@buffalo.edu.

Abstract

In this paper, we study the Principal Component Analysis (PCA) problem under the (distributed) non-interactive local differential privacy model. For the low dimensional case, we show the optimal ratefor the private minimax risk of the k-dimensional PCA using the squared subspace distance as the measurement. For the high dimensional row sparse case, we first give a lower bound on the private minimax risk, . Then we provide an efficient algorithm to achieve a near optimal upper bound. Experiments on both synthetic and real world datasets confirm the theoretical guarantees of our algorithms.

1 Introduction

Principal Component Analysis (PCA) is a fundamental technique for dimension reduction in statistics, machine learning, and signal processing. As of today, it remains as one of the most commonly used tools in applications, especially in social sciences [Costello and Osborne, 2005], financial econometrics [Aït-Sahalia and Xiu, 2017], medicine [Barber *et al.*, 1975], and genomics [Lu and Xu, 2013].

With the rapid development of information technologies, big data now ubiquitously exist in our daily life, which need to be analyzed (or learned) statistically by methods like regression and PCA. However, due to the presence of sensitive data (especially those in social science, biomedicine and genomics) and their distributed nature, such data are extremely difficult to aggregate and learn from. Consider a case where health records are scattered across multiple hospitals (or even countries), it is challenging to process the whole dataset in a central server due to privacy and ownership concerns. A better solution is to use some differentially private mechanisms to conduct the aggregation and learning tasks. Differential Privacy (DP) (Dwork et al., 2006b) is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

*The research of this work was supported in part by NSF through grant CCF-1716400.

†Contact Author

Currently, there are mainly two user models available for differential privacy: the central model and the local one. In the central model, data are managed by a trusted central entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release. A classical application of this model is the one of collecting census data. In the local model instead, each individual manages his/her proper data and discloses them to a server through some differentially private mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. A classical example of this model is the one aiming at collecting statistics from user devices like in the case of Google's Chrome browser [Erlingsson *et al.*, 2014], and Apple's iOS-10 [Near, 2018].

In the local model, two basic types of protocols are often used: interactive and non-interactive. [Smith *et al.*, 2017] have recently investigated the power of non-interactive differentially private protocols. This type of protocols is more natural for the classical use cases of the local model: both projects from Google and Apple use the non-interactive model. Moreover, implementing efficient interactive protocols in such applications is more difficult due to the latency of the network. Despite being used in industry, the local model has been much less studied than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

In this paper, we study PCA under the non-interactive local differential privacy model and aim to answer the following main question.

What are the limitations and the (near) optimal algorithms of PCA under the non-interactive local differential privacy model?

We summarize our main contributions as follows:

1. We first study the k-subspace PCA problem in the low dimensional setting and show that the minimax risk (measured by the squared subspace distance) under ϵ non-interactive local differential privacy (LDP) is lower bounded by $\Omega(\frac{\lambda_1\lambda_{k+1}pk}{(\lambda_k-\lambda_{k+1})^2n\epsilon^2})$, where p is the dimensionality of the data and n is the number of data records. Moreover, we prove that the term $\Omega(\frac{pk}{n\epsilon^2})$ is optimal.
2. An undesirable issue of the above result is that the er-

ror bound could be too large in high dimensions (*i.e.,* $p \gg n$). In such scenarios, a natural approach is to impose some additional structural constraints on the leading eigenvectors. A commonly used constraint is to assume that the leading eigenvectors are row sparse, which is refereed as sparse PCA in the literature and has been studied intensively in recent years [Vu *et al.*, 2013b; Cai *et al.*, 2013; Vu *et al.*, 2013a]. Thus, for the high dimensional case, we consider the sparse PCA under the non-interactive local model and show that the private minimax risk (measured by the squared subspace distance) is lower bounded by $\Omega(\frac{\lambda_1\lambda_{k+1}}{(\lambda_k-\lambda_{k+1})^2}\frac{ks\log p}{n\epsilon^2})$, where λ_1, λ_k and λ_{k+1} are eigenvalues and s is the sparsity parameter of the eigenvectors. We also give an algorithm to achieve a near optimal upper bound of $O\big(\frac{\lambda_1^2}{(\lambda_k-\lambda_{k+1})^2}\frac{s^2\log p}{n\epsilon^2}\big)$.

2 Related Work

There is a vast number of papers studying PCA under differential privacy, starting from the SULQ framework [Blum *et al.*, 2005], [Dwork *et al.*, 2014; Chaudhuri *et al.*, 2013; Jiang *et al.*, 2016; Gonem and Gilad-Bachrach, 2018; Ge *et al.*, 2018; Balcan *et al.*, 2016]. We compare only those private PCA results in distributed settings.

For the low dimensional case, Balcan *et al.* [Balcan *et al.*, 2016] studied the private PCA problem under the interactive local differential privacy model and introduced an approach based on the noisy power method. They showed an upper bound which is suitable for general settings, while ours is mainly for statistical settings. It is worth pointing out that the output in [Balcan *et al.*, 2016] is only an $O(k)$-dimensional subspace, instead of an exact k-dimensional subspace; thus their result is incomparable with ours. Moreover, we provide, in this paper, a lower bound on the ϵ non-interactive private minimax risk.

For the private high dimensional sparse PCA, the work most closely related to ours is the one by Ge *et al.* [Ge *et al.*, 2018]. The authors in this paper proposed a noisy iterative hard thresholding power method, which is an interactive LDP algorithm and proved an upper bound of $O\big(\frac{\lambda_1\lambda_k}{(\lambda_k-\lambda_{k+1})^2}\frac{s(k+\log p)}{n(1-\rho^{\frac{1}{4}})}\big)$ for their method, where ρ is a parameter related to ϵ. Specifically, they showed that there exists some 'Privacy Free Region'. However, several things need to be pointed out. Firstly, our method is for general $\epsilon \in (0, 1]$ and non-interactive settings, while Ge *et al.* considered the interactive setting with more restricted ϵ. Secondly, the assumptions in our paper are less strict than the ones in [Ge *et al.*, 2018]. Finally, we provide a lower bound on the private minimax risk.

The optimal procedure in our paper is based on perturbing the covariance by Gaussian matrices, which has been studied in [Dwork *et al.*, 2014]. However, there are some major differences; firstly, we show the optimality of our algorithm under the non-interactive local model using subspace distance as the measurement, while [Dwork *et al.*, 2014] showed the optimality under the (ϵ, δ) central model using variance as the measurement. It is notable that in [Dwork *et al.*, 2014] the authors also provided an upper bound on the subspace distance. However, the lower bound is still unknown. Secondly, while the optimal algorithm for the low dimensional case is quite similar, we extend it to the high dimensional case. The optimal procedure in the high dimensional sparse case is quite different from that in [Dwork *et al.*, 2014]. Thirdly, in this paper,since we focus the statistical setting while [Dwork *et al.*, 2014] considered the general setting, the upper bound results are incomparable.

3 Preliminaries

3.1 Classical Minimax Risk

Since all of our lower bounds are in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing the locally private version.

Let $\mathcal{P}$ be a class of distributions over a data universe $\mathcal{X}$. For each distribution $P \in \mathcal{P}$, there is a deterministic function $\theta(P) \in \Theta$, where Θ is the parameter space. Let $\rho : \Theta\times\Theta :\mapsto \mathbb{R}_+$ be a semi-metric function on the space Θ and $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (in this paper, we assume that ρ is the subspace distance and $\Phi(x) = x^2$ unless specified otherwise).

We further assume that $\{X_i\}_{i=1}^n$ are n i.i.d observations drawn according to some distribution $P \in \mathcal{P}$, and $\hat{\theta} : \mathcal{X}^n \mapsto \Theta$ is some estimator. Then, the minimax risk in metric $\Phi\circ\rho$ is defined by the following saddle point problem:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi\circ\rho) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[\Phi(\rho(\hat{\theta}(X_1, \cdots, X_n), \theta(P))],$$

where the supremum is taken over distributions $P \in \mathcal{P}$ and the infimum over all estimators $\hat{\theta}$.

3.2 Local Differential Privacy and Private Minimax Risk

Since we consider the non-interactive local model in this paper, we will follow the definitions in [Duchi *et al.*, 2013].

We assume that $\{Z_i\}_{i=1}^n$ are the private observations transformed from $\{X_i\}_{i=1}^n$ through some privacy mechanisms. When Z_i depends only on X_i, the mechanism is called non-interactive and in this case we have a simpler form for the conditional distributions $Q_i(Z_i \mid X_i = x_i)$. We now define local differential privacy by restricting the conditional distribution.

Definition 1 ([Duchi *et al.*, 2013])**.** For a given privacy parameter $\epsilon, \delta > 0$, we say that the random variable Z_i is an (ϵ, δ) non-interactively locally differentially private (LDP) view of X_i if for any $x_i, x_i' \in \mathcal{X}$:

$$Q_i(Z_i \in S \mid X_i = x_i) \le e^\epsilon Q_i(Z_i \in S \mid X_i = x_i') + \delta.$$

When $\delta = 0$, we call it ϵ non-interactively LDP view. We say that the privacy mechanism $Q = \{Q_i\}_{i=1}^n$ is (ϵ, δ) (ϵ) non-interactively locally differentially private (LDP) if each Z_i is an (ϵ, δ) (ϵ) non-interactively LDP view.

For a given privacy parameter $\epsilon > 0$, let $\mathcal{Q}_\epsilon$ be the set of conditional distributions that have the ϵ-LDP property. For a

given set of samples $\{X_i\}_{i=1}^n$, let $\{Z_i\}_{i=1}^n$ be the set of observations produced by any distribution $Q \in \mathcal{Q}_\epsilon$. Then, our estimator will be based on $\{Z_i\}_{i=1}^n$, that is, $\hat{\theta}(Z_1, \cdots, Z_n)$. This yields a modified version of the minimax risk:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi\circ\rho, Q) = \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[\Phi(\rho(\hat{\theta}(Z_1, \cdots, Z_n), \theta(p))].$$

From the above definition, it is natural to seek the mechanism $Q \in \mathcal{Q}_\epsilon$ that has the smallest value for the minimax risk. This allows us to define functions that characterize the optimal rate of estimation in terms of privacy parameter ϵ.

Definition 2. Given a family of distributions $\theta(\mathcal{P})$ and a privacy parameter $\epsilon > 0$, the ϵ non-interactive private minimax risk in the metric $\Phi\circ\rho$ is:

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi\circ\rho, \epsilon) := \inf_{Q\in\mathcal{Q}_\epsilon} \mathcal{M}_n(\theta(\mathcal{P}), \Phi\circ\rho, Q),$$

where $\mathcal{Q}_\epsilon$ is the set of all ϵ non-interactively locally differentially private mechanisms.

3.3 Locally Private k-dimensional PCA

Let $X \in \mathbb{R}^p$ a random vector with mean 0 and covariance matrix Σ. k-dimensional PCA is to find a k dimensional subspace that optimizes the following problem:

$$\min \mathbb{E}\|(I_p - \Pi_{\mathcal{G}})X\|_2^2, \text{ s.t. } \mathcal{G} \in \mathbb{G}_{p,k},$$

where $\mathbb{G}_{p,k}$ is the Grassmann manifold of k-dimensional subspaces of $\mathbb{R}^p$, and $\Pi_{\mathcal{G}}$ is the projection of $\mathcal{G}$. There always exists at least one solution; consider $\Sigma = \sum_{j=1}^p \lambda_j v_j v_j^T$, where $\lambda_1 \geq \lambda_2 \geq \cdots, \lambda_p \geq 0$ are the eigenvalues of Σ and $v_1, v_2, \cdots, v_p \in \mathbb{R}^p$ are the corresponding eigenvectors. If $\lambda_k \geq \lambda_{k+1}$, then the k-dimensional principal subspace of Σ, *i.e.* the subspace $\mathcal{S}$ spanned by $v_1, \cdots, v_k$ solves the above optimization problem, where the orthogonal projector of $\mathcal{S}$ is given by $\Pi_{\mathcal{S}} = V_k V_k^T$, where $V_k = [v_1, \cdots, v_k] \in \mathbb{V}_{p,k}$, $\mathbb{V}_{p,k}$ is the set of all $p \times k$ orthogonal matrices. For simplicity we denote $\mathcal{S} = \text{col}(V_k)$, where $\text{col}(M)$ denotes the subspace spanned by the columns vectors of M.

In practice, Σ is unknown, and the only thing that we have is the set of observation data records $\{X_1, \cdots, X_n\}$, which are i.i.d sampled from X. Thus, the problem of (non-interactively) locally differentially private PCA is to find a k-dimensional subspace $\mathcal{S}^{\text{priv}}$ which is close to $\mathcal{S}$, where the algorithm that outputs $\mathcal{S}^{\text{priv}}$ must be ϵ (non-interactively) locally differentially private. After obtaining a private estimator $\mathcal{S}^{\text{priv}}$, there are multiple ways to measure the success, such as variance guarantee [Dwork *et al.*, 2014], low rank approximation error [Kapralov and Talwar, 2013], etc. In this paper, we will use the subspace distance as the measurement [Dwork *et al.*, 2014; Ge *et al.*, 2018].

Let $\mathcal{S}$ and $\mathcal{S}'$ be two k-dimensional subspaces in $\mathbb{R}^p$. Also denote by E and F, respectively, the orthogonal matrix corresponds to $\mathcal{S}$ and $\mathcal{S}'$. That is, $E = VV^T$ and $F = WW^T$ for some orthogonal matrices $V \in \mathbb{V}_{p,k}$ and $W \in \mathbb{V}_{p,k}$. Then, the squared subspace distance between $\mathcal{S}$ and $\mathcal{S}'$ is defined by the following [Stewart, 1990]:

$$\|\sin\Theta(\mathcal{S}, \mathcal{S}')\|_F^2 = \|E - F\|_F^2 = \frac{1}{2}\|VV^T - WW^T\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenious norm. For simplicity, we will overload notation and write $\sin\Theta(\mathcal{S}, \mathcal{S}') = \sin\Theta(V, W)$.

4 Low Dimensional Case

In this section, we focus on the general case and always assume $n \geq p$. We first derive a lower bound of the ϵ non-interactive private minimax risk using the squared subspace distance as the measurement. By the definition of the ϵ-private minimax risk, it is important to select an appropriate class of distributions.

4.1 Class of Distributions

We assume that the observations $\{X_i\}_{i=1}^n$ satisfies: $X = \Sigma^{\frac{1}{2}} Z$, where $Z \in \mathbb{R}^p$ is a random vector satisfying equations $\mathbb{E}Z = 0$ and $\text{Var}(Z) = I_p$. Also, Z is sub-Gaussian and $\|Z\|_{\psi_2} \leq 1$ where

$$\|Z\|_{\psi_2} := \sup_{v:\|v\|_2\leq 1} \inf\{C > 0, \mathbb{E}\exp|\frac{< Z, v >}{C}|^2 \leq 2\},$$

which means that all the one-dimensional marginals of X_i have sub-Gaussian tails. We need to note that this assumption on X is commonly used in many papers on PCA in statistical settings, such as [Vu *et al.*, 2013b; Ge *et al.*, 2018].

In the study of private PCA, it is always assumed that the ℓ_2 norm of each X_i is bounded by 1, as in [Dwork *et al.*, 2014][Ge *et al.*, 2018]. For convenience, we relax this assumption in the following way; for the random vector $X \in \mathbb{R}^p$, we assume that $\|X\|_2 \leq 1$ with a probability at least $1 - e^{-\Omega(p)}$.

Next, we give assumptions on the population covariance matrix Σ. Firstly, we assume that for the target k-dimensional subspace, $\lambda_k - \lambda_{k+1} > 0$ so that the principal subspace is well defined. Next, we define the effective noise variance, which is proposed in [Vu *et al.*, 2013b] and [Cai *et al.*, 2013]:

$$\sigma_k^2(\lambda_1, \lambda_2, \cdots, \lambda_p) := \frac{\lambda_1\lambda_{k+1}}{(\lambda_k - \lambda_{k+1})}. \tag{1}$$

For a given constant $\sigma^2 > 0$, we assume that $\sigma_k^2 \leq \sigma^2$. Also, we denote the collection of distributions by $\mathcal{P}(k, \sigma^2)$.

4.2 Main Results

The next theorem shows a lower bound of ϵ non-interactive private minimax risk under squared subspace distance.

Theorem 1. Let $\{X_i\}_{i=1}^n$ be samples from $P \in \mathcal{P}(k, \sigma^2)$. If $\frac{p}{4} \leq k \leq \frac{3p}{4}$, $\epsilon \in (0, \frac{1}{2}]$ and $n \geq \Omega\big(\frac{1}{\epsilon^2}\frac{\lambda_1\lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}\min\{k, p-k\}\big)$, then the ϵ non-interactive private minimax risk in the metric of squared subspace distance satisfies:

$$\mathcal{M}_n^{\text{Nint}}(\mathcal{S}(\mathcal{P}(k, \sigma^2)), \Phi\circ\rho, \epsilon) \geq \Omega\Big(\frac{\lambda_1\lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}\frac{kp}{n\epsilon^2}\Big).$$

Remark 1. We note that for the non-private case, the minimax risk is lower bounded by $\Omega\big(\frac{\lambda_1\lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}\frac{kp}{n}\big)$ [Cai *et al.*, 2013]. Thus, in this case, the impact of the local differential privacy is to change the number of efficient sample from n to $n\epsilon^2$. However, the collection of the considered distributions needs another assumption which says that $\|X\|_2$ is bounded by 1 with high probability. This is not necessary in the non-private case [Cai *et al.*, 2013], but needed in ours for showing the upper bound.

We also note that in the central differential privacy model, [Dwork *et al.*, 2014] showed that the lower bound of the k-dimensional PCA is $\tilde{\Omega}(\frac{kp\log(\frac{1}{\delta})}{n^2\epsilon^2})$ for (ϵ, δ)-differential privacy. However, this lower bound is measured by the variance of $X = (X_1^T, X_2^T, \cdots, X_n^T)^T \in \mathbb{R}^{n\times p}$, not the squared subspace distance used in this paper. Although [Dwork *et al.*, 2014] gave an upper bound of $O(\frac{kp\log(1/\delta)}{(\lambda_k^2-\lambda_{k+1}^2)n^2\epsilon^2})$ in the general setting using the squared subspace distance as measurement, it is still unknown whether the bound is optimal. Also, their lower bound omits the parameters related to the eigenvalues. For the ϵ differential privacy in the central model, [Chaudhuri *et al.*, 2013] showed that the lower bound is $\Omega(\frac{p^2}{n^2\epsilon^2(\lambda_1-\lambda_2)^2})$ in the special case of $k = 1$. However, it is still unknown for the general case of k. Thus, from the above discussion, we can see that the lower bound of ϵ non-interactively locally differentially private PCA is similar to the (ϵ, δ) differentially private PCA in the central model.

One of the main questions is whether the lower bound in Theorem 1 is tight. In the following, we show that the term $\Omega(\frac{pk}{n\epsilon^2})$ is tight. By our definition of the parameter space, we know that for any $X \sim P \in \mathcal{P}(\sigma^2, k)$, $\|X\|_2 \le 1$ with high probability. Thus, we always assume that the event of each $\|X_i\|_2 \le 1$ holds. Note that this assumption also appears in [Ge *et al.*, 2018; Dwork *et al.*, 2014; Balcan *et al.*, 2016]. The idea is the same as in [Dwork *et al.*, 2014], where each X_i perturbs its covariance and aggregates the noisy version of covariance, see Algorithm 1 for details.

Theorem 2. For any $\epsilon, \delta > 0$, Algorithm 1 is (ϵ, δ) (non-interactively) locally differentially private. Furthermore, with probability at least $1 - e^{-C_1} - \frac{1}{p^{C_2}}$, the output satisfies:

$$\|\sin\Theta(\tilde{V}_k, V_k)\|_F^2 \le O\Big(\frac{\lambda_1^2 kp\log(1/\delta)}{(\lambda_k - \lambda_{k+1})^2 n\epsilon^2}\Big), \qquad (2)$$

where C_1, C_2 are some universal constants.

Algorithm 1 Local Gaussian Mechanism

Input: data records $\{X_i\}_{i=1}^n \sim P^n$ for $P \in \mathcal{P}(\sigma^2, k)$, and for $i \in [n]$, $\|X\|_2 \le 1$. ϵ, δ are the privacy parameters.

1: **for** Each $i \in [n]$ **do**
2: Denote $\tilde{X}_i\tilde{X}_i^T = X_iX_i^T + Z_i$, where $Z_i \in \mathbb{R}^{p\times p}$ is a symmetric matrix where the upper triangle (including the diagonal) is i.i.d samples from $\mathcal{N}(0, \sigma_1^2)$; here $\sigma_1^2 = \frac{2\ln(1.25/\delta)}{\epsilon^2}$, and each lower triangle entry is copied from its upper triangle counterpart.
3: **end for**
4: Compute $\tilde{S} = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^T$.
5: Output $\mathrm{col}(\tilde{V}_k)$ where $\tilde{V}_k \in \mathbb{R}^{p\times k}$ is the principal rank k subspace of $\tilde{S}$.

In Theorem 7 of [Dwork *et al.*, 2014], the authors provided a similar upper bound for the (ϵ, δ)-differential privacy in the central model. However, they need to assume that the eigenvalues satisfy the condition $\lambda_k^2 - \lambda_{k+1}^2 = \omega(\sqrt{p})$, which is not needed in our Theorem 2.

From Theorems 1 and 2, we can see that there is still a gap of $O(\frac{\lambda_1}{\lambda_{k+1}})$ between the lower and upper bounds. We leave it as an open problem to determine whether these bounds are tight or not.

5 High Dimensional Sparse Case

From Theorem 1, we can see that for the high dimensional case, *i.e.* $p \gg n$, the bound in (2) becomes trivial. Thus, to avoid this issue, we need some additional assumption on the parameter space. One of the commonly used assumption is sparsity. There are many definitions of sparsity on PCA and we use the row sparsity in this paper, which has also been studied in [Vu *et al.*, 2013b; Cai *et al.*, 2013; Ge *et al.*, 2018].

We first define the $(2, 0)$-norm of a $p \times k$ matrix A as the usual ℓ_2 norm of the vector of row-wise ℓ_0 norms of A:

$$\|A\|_{2,0} := \|(\|a_{1*}\|_2, \|a_{2*}\|_2, \cdots, \|a_{p*}\|_2)\|_0, \qquad (3)$$

where a_{j*} denotes the j-th row of A. Note that $\|\cdot\|_{2,0}$ is coordinate independent, *i.e.* $\|AO\|_{2,0} = \|A\|_{2,0}$ for any orthogonal matrix $O \in \mathbb{R}^{k\times k}$. We define the row sparse space as follows.

Definition 3. Let s be the sparsity level parameter satisfying the condition of $k \le s \le p$. The s-(row) sparse subspace is defined as follows

$$\mathcal{M}_0(s) = \{\mathrm{Col}(U), U \in \mathbb{R}^{p\times k} \text{ and orthogonal }, \|U\|_{2,0} \le s\}.$$

We define our parameter space, $\mathcal{P}(s, k, \sigma^2)$, to be the same as in the previous section with an additional condition that $S \in \mathcal{M}_0(s)$, where S is the k-dimensional principal subspace of covariance matrix Σ.

Below, we will first derive a lower bound of the non-interactive locally differentially private PCA in the high dimensional sparse case.

Theorem 3. Let $\{X_i\}_{i=1}^n$ be the observations sampled from a distribution $P \in \mathcal{P}(s, k, \sigma^2)$. If the privacy parameter $\epsilon \in (0, \frac{1}{2}]$, $n \ge C(s-k)\frac{\sigma^2(k+\log p)}{\epsilon^2}$ for a universal constant $C > 0$, and $2k \le s - k \le p - k$, then for all $k \in [p]$ satisfying the condition of $k \le p - 4$, the ϵ non-interactive private minimax risk in the metric of squared subspace distance satisfies the following

$$\mathcal{M}_n^{\mathrm{Nint}}(\mathcal{S}(\mathcal{P}(s, k, \sigma^2), \epsilon) \ge \Omega\Big(\frac{\lambda_1\lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}\frac{s(k+\log p)}{n\epsilon^2}\Big).$$

Note that in the non-private case, the optimal minimax risk is $O\big(\frac{\lambda_1\lambda_{k+1}}{(\lambda_k-\lambda_{k+1})^2}\frac{s(k+\log p)}{n}\big)$. Thus, same as in the low dimensional case, the impact of the privacy constraint is to change the efficient samples from n to $n\epsilon^2$.

Next, we consider the upper bound. In the non-private case, the optimal procedure is to solve the following NP-hard optimization problem [Vu *et al.*, 2013b]:

$$\max < S, UU^T >$$
$$\text{subject to } U^TU = I_k, U \in \mathbb{R}^{p\times k} \text{ and } \|U\|_{2,0} \le s, \qquad (4)$$

where S is the empirical covariance matrix. Our upper bound is based on (4). However, instead of solving (4) on the perturbed version of the empirical covariance matrix, we perturb the covariance matrix and solve the following optimization problem on the convex hull of the constraints in (4), that is:

$$\hat{X} = \arg\max < \tilde{S}, X > -\lambda \|X\|_{1,1} \quad (5)$$
$$\text{subject to } X \in \mathcal{F}^k := \{X : 0 \preceq X \preceq I \text{ and } \mathrm{Tr}(X) = k\},$$

where $< S, X >= \mathrm{Tr}(SX^T)$. Note that the constraints in (5), which is called Fantope [Bhatia, 2013][Vu *et al.*, 2013a], is the convex hull of the constrains in (4). Also, since the constraints in (5) only guarantees that the rank of the output is $\geq k$, the output $\hat{X}$ needs not to be a matrix with exact rank of k. Thus, in order to obtain a proper k-dimensional subspace, we just output the k-PCA of $\hat{X}$.

Algorithm 2 Local Gaussian Mechanism-High Dimension

Input: data records $\{X_i\}_{i=1}^n \sim P^n$ for $P \in \mathcal{P}(s, \sigma^2, k)$, and for $i \in [n]$, $\|X\|_2 \leq 1$. ϵ, δ are privacy parameters. $\rho >$ is a constant.

1: **for** Each $i \in [n]$ **do**
2: Denote $\tilde{X}_i \tilde{X}_i^T = X_i X_i^T + Z_i$, where $Z_i \in \mathbb{R}^{p \times p}$ is a symmetric matrix where the upper triangle (including the diagonal) is i.i.d samples from $\mathcal{N}(0, \sigma_1^2)$; here $\sigma_1^2 = \frac{2\ln(1.25/\delta)}{\epsilon^2}$, and each lower triangle entry is copied from its upper triangle counterpart.
3: **end for**
4: Compute $\tilde{S} = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T$.
5: Get the optimal solution $\hat{X}$ in (5) or do as the followings
6: Setting $Y^{(0)} = 0, U^{(0)} = 0$
7: **for** $t = 1, 2, \cdots$ **do**
8: $X^{(t+1)} = \mathcal{P}_{\mathcal{F}^k}(Y^{(t)} - U^{(t)} + \frac{\tilde{S}}{\rho})$
9: $Y^{(t+1)} = \mathcal{S}_{\lambda/\rho}(X^{(t+1)} + U^{(t)})$ where $\mathcal{S}$ is the entry-wise soft thresholding operator defined as $\mathcal{S}_{\lambda/\rho}(x) = \mathrm{sign}(x)\max(|x| - \lambda/\rho, 0)$.
10: $U^{(t+1)} = U^{(t)} + X^{(t+1)} - Y^{(t+1)}$
11: Return $Y^{(t)}$
12: **end for**
13: Let k-dimensional principal component of $\hat{X}$ or $Y^{(t)}$ be $\tilde{V}_k$, output $\hat{\mathcal{S}} = \mathrm{col}(\tilde{V}_k)$.

Theorem 4. For any given $0 < \epsilon, \delta < 1$, if $\{X_i\}_{i=1}^n \sim P^n$ for $P \in \mathcal{P}(s, \sigma^2, k)$ and $\|X\|_2 \leq 1$ for all $i \in [n]$, then the solution to the optimization problem (5) is (ϵ, δ) non-interactive locally differentially private. Moreover, if let $\hat{V}_k$ denote the k-dimensional principal component subspace of $\hat{X}$ and set $\lambda \leq O(\lambda_1 \sqrt{\frac{\log p}{n\epsilon^2}})$, then with probability at least $1 - \frac{2}{p^2} - \frac{1}{p^c}$, the following holds

$$\|\sin\Theta(\hat{V}_k, V_k)\|_F^2 \leq O\Big(\frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s^2 \log p}{n\epsilon^2}\Big),$$

where c is a universal constant.

Since the optimization problem (5) is convex, we can follow the approach in [Vu *et al.*, 2013a] to solve it by using ADMM method (see Algorithm 2 for the details).

Comparing with the lower bound of the private minimax risk in Theorem 3, we can see that the bound in Theorem 4 is roughly larger than the optimal rate by a factor of $O(\frac{\lambda_1}{\lambda_{k+1}}\frac{s}{k})$. This means that the upper bound is only near optimal [Vu *et al.*, 2013a]. A remaining open problem is to determine whether it is possible to get a tighter upper bound that does not contain the term of $\frac{s}{k}$ in the gap.

6 Experiments

Dataset	Size	s	Error
cancer RNA-Seq	(801, 20531)	10	3.162
		20	3.381
		40	3.668
Leukemia	(72, 7128)	10	3.162
		20	3.435
		40	3.701
Colon cancer	(60, 2000)	10	2.449
		20	3.058
		40	3.228
isolet5	(1559, 617)	10	1.441
		20	2.023
		40	2.508
lung	(203, 3312)	10	2.858
		20	3.464
		40	3.901
NIPS	(11463, 5811)	10	3.643
		20	3.881
		40	4.472

Table 1: Results with different sparsity s for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions k is set to be $k = 10$ and $\epsilon = 2$.

6.1 Low Dimensional Case

For synthetic datasets, we generate the data samples $\{X_i\}_{i=1}^n$ independently from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma = \frac{\lambda}{5p(\lambda+1))} VV^T + \frac{1}{5p(\lambda+1)} I_p$ for $V \in \mathbb{V}_{p,k}$. It can be shown that $\|X_i\|_2 \leq 1\ \forall i \in [n]$ with high probability (see supplemental material). We choose $n = 10^5$, $p = 40$, $k = \{5, 10, 15, 20\}$, $\epsilon = 0.5$, $\delta = 10^{-4}$, and $\lambda = 1$. For real world datasets, we run Algorithm 1 on Covertype and Buzz datasets [Dheeru and Karra Taniskidou, 2017] with normalized rows for each dataset. The error is measured by the subspace distance $\|\hat{V}_k \hat{V}_k^T - V_k V_k^T\|_F$. For each experiment, we repeat 20 times and take the average as the final result.

Figure 1 is the result for the synthetic datasets. Figure indicates that 1) the error deceases as the sample size increases or ϵ increases (*i.e., becomes less private*); 2) the error increases as the dimensionality p increases or the dimensionality k of the target subspace increases.

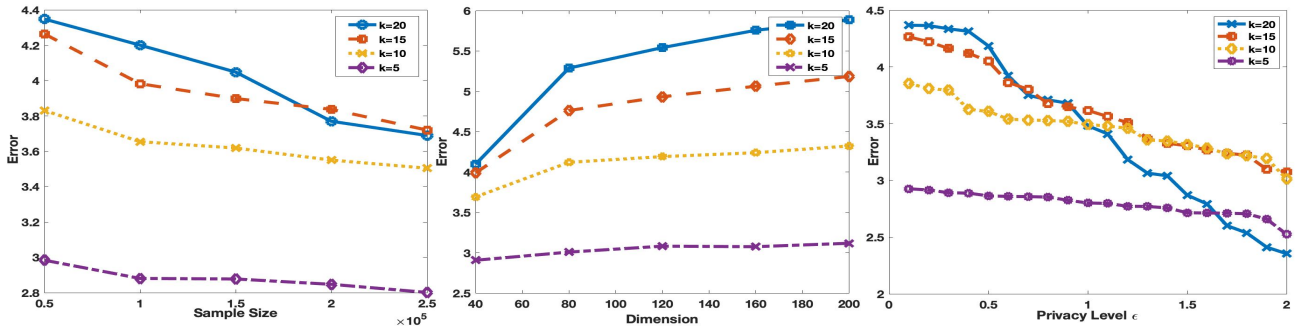

Figure 1: LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions k over sample size n with fixed $\epsilon = 2$ and $p = 400$. The middle one is for different dimensions with fixed $n = 2000$ and $\epsilon = 2$. The right one is for different level of privacy with fixed $n = 2000$ and $p = 400$.

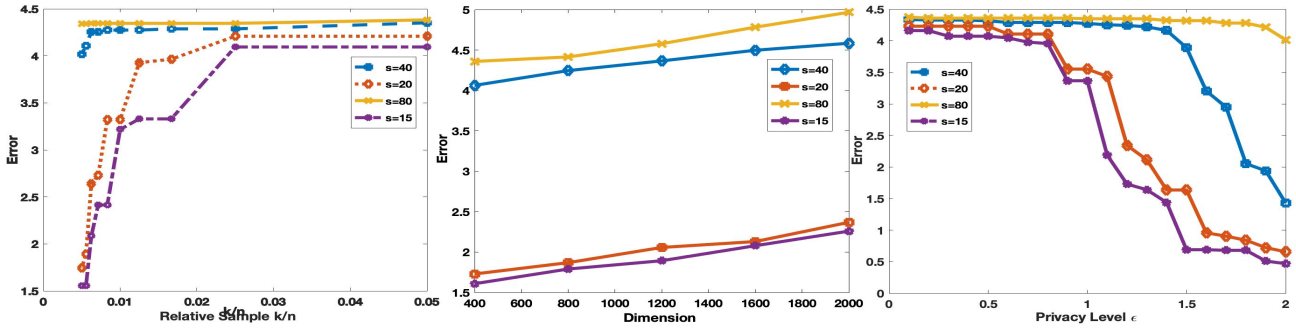

Figure 2: LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions k over sample size n with fixed $\epsilon = 2$ and $p = 400$. The middle one is for different dimensions with fixed $n = 2000$ and $\epsilon = 2$. The right one is for different level of privacy with fixed $n = 2000$ and $p = 400$.

6.2 High Dimensional Case

For the high dimensional case, we consider the same distributions as in the low dimensional case and generate the target subspace V in the following way. For a given sparsity parameter s, we first generate a random orthogonal matrix $\tilde{V} \in \mathbb{R}^{s\times k}$, then pad it with rows of zeros, and finally randomly permute the matrix. We set $k = 10, n = 2000, p = 400$, $s = \{15, 20, 40, 80\}$ and $\epsilon = 1$.

Besides the synthetic datasets, we also test our algorithm on some real world datasets in [Dheeru and Karra Taniskidou, 2017] and [Li *et al.*, 2017]. We can see that 1) as the term of $\frac{k}{n}$ increases (n decreases), the error increases accordingly; 2) the error slightly increases when the dimensionality p increases, which is due to the fact that the upper bound in Theorem 4 depends only logarithmically on p (*i.e.,* $\log p$); 3) the error decreases when ϵ increases. Table 1 and 2 show the results of the error with different sparsity and privacy, respectively.

Dataset	Size	ϵ	Error
cancer RNA-Seq	(801, 20531)	1	3.559
		0.5	3.790
		0.1	3.967
Leukemia	(72, 7128)	1	4.375
		0.5	4.403
		0.1	4.518
Colon cancer	(60, 2000)	1	3.013
		0.5	4.237
		0.1	4.310
isolet5	(1559, 617)	1	2.884
		0.5	3.405
		0.1	3.896
lung	(203, 3312)	1	4.042
		0.5	4.275
		0.1	4.362
NIPS	(11463, 5811)	1	4.006
		0.5	4.052
		0.1	4.472

Table 2: Results with different privacy levels ϵ for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions k is set to be $k = 10$ and $s = 20$.

References

[Aït-Sahalia and Xiu, 2017] Yacine Aït-Sahalia and Dacheng Xiu. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399, 2017.

[Balcan *et al.*, 2016] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309, 2016.

[Barber *et al.*, 1975] DC Barber, PJ Howlett, and RC Smart. Principal component analysis in medical research. *Journal of Applied Statistics*, 2(1):39–43, 1975.

[Bhatia, 2013] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[Blum *et al.*, 2005] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138. ACM, 2005.

[Cai *et al.*, 2013] T Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[Chaudhuri *et al.*, 2013] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14(1):2905–2943, 2013.

[Costello and Osborne, 2005] Anna B Costello and Jason W Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9, 2005.

[Dheeru and Karra Taniskidou, 2017] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[Duchi *et al.*, 2013] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

[Dwork *et al.*, 2014] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20. ACM, 2014.

[Erlingsson *et al.*, 2014] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, pages 1054–1067. ACM, 2014.

[Ge *et al.*, 2018] Jason Ge, Zhaoran Wang, Mengdi Wang, and Han Liu. Minimax-optimal privacy-preserving sparse pca in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, pages 1589–1598, 2018.

[Gonem and Gilad-Bachrach, 2018] Alon Gonem and Ram Gilad-Bachrach. Smooth sensitivity based approach for differentially private pca. In *Algorithmic Learning Theory*, pages 438–450, 2018.

[Jiang *et al.*, 2016] Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. In *AAAI*, pages 1730–1736, 2016.

[Kapralov and Talwar, 2013] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1395–1414. SIAM, 2013.

[Li *et al.*, 2017] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2017.

[Lu and Xu, 2013] Dongsheng Lu and Shuhua Xu. Principal component analysis reveals the 1000 genomes project does not sufficiently cover the human genetic diversity in asia. *Frontiers in genetics*, 4:127, 2013.

[Near, 2018] Joe Near. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, Santa Clara, CA, 2018. USENIX Association.

[Smith *et al.*, 2017] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 58–77. IEEE, 2017.

[Stewart, 1990] G. W. Stewart. Matrix perturbation theory, 1990.

[Vu *et al.*, 2013a] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in neural information processing systems*, pages 2670–2678, 2013.

[Vu *et al.*, 2013b] Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.